massachusetts institute of technology — computer science and artificial intelligence laboratory

Risk Bounds for Mixture Density Estimation

Alexander Rakhlin, Dmitry Panchenko and Sayan Mukherjee

AI Memo 2004-001 January 2004
CBCL Memo 233

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JAN 2004		**00-01-2004 to 00-01-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Risk Bounds for Mixture Density Estimation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Artificial Intelligence Laboratory,77 Massachusetts Avenue,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**12**	

Abstract

In this paper we focus on the problem of estimating a bounded density using a finite combination of densities from a given class. We consider the Maximum Likelihood Procedure (MLE) and the greedy procedure described by Li and Barron [6, 7]. Approximation and estimation bounds are given for the above methods. We extend and improve upon the estimation results of Li and Barron, and in particular prove an $O(\frac{1}{\sqrt{n}})$ bound on the estimation error which does not depend on the number of densities in the estimated combination.

This report describes research done at the Center for Biological& Computational Learning, which is in the Department of Brain & Cognitive Sciences at MIT and which is affiliated with the McGovern Institute of Brain Research and with the Artificial Intelligence Laboratory.
This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, and National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506.
Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), DaimlerChrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., The Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd, and the SLOAN Foundation.

1 Introduction

In the density estimation problem, we are given n i.i.d. samples $S = \{x_1, ..., x_n\}$ drawn from an unknown density f. The goal is to estimate this density from the given data. We consider the Maximum Likelihood Procedure (MLE) and the greedy procedure described by Li and Barron [6, 7] and prove estimation bounds for these procedures. Rates of convergence for density estimation were studied in [3, 10, 11, 13]. For neural networks and projection pursuit, approximation and estimation bounds can be found in [1, 2, 4, 9].

To evaluate the accuracy of the density estimate we need a notion of distance. Kullback-Leibler (KL) divergence and Hellinger distance are the most commonly used. Li and Barron [6, 7] give final bounds in terms of KL-divergence, and since our paper extends and improves upon their results, we will be using this notion of distance as well. The KL-divergence between two distributions is defined as

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx = \mathbb{E} \log \frac{f}{g}.$$

The expectation here is assumed to be with respect to x, which comes from a distribution with the density $f(x)$.

Consider a parametric family of probability density functions $\mathcal{H} = \{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\}$. The class of k-component mixtures f_k is defined as

$$f_k \in \mathcal{C}_k = \text{conv}_k(\mathcal{H}) = \left\{ f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \theta_i \in \Theta \right\}.$$

Approximation results will depend on the following class of continuous convex combinations (with respect to all measures P on Θ)

$$\mathcal{C} = \text{conv}(\mathcal{H}) = \left\{ f : f(x) = \int_\Theta \phi_\theta(x) P(d\theta) \right\}.$$

The approximation bound of Li and Barron [6, 7] states that for any f, there exists an $f_k \in \mathcal{C}_k$, such that

$$D(f\|f_k) \leq D(f\|\mathcal{C}) + \frac{c_{f,P}^2 \gamma}{k}, \tag{1}$$

where $c_{f,P}$ and γ are constants and $D(f\|\mathcal{C}) = \inf_{g \in \mathcal{C}} D(f\|g)$. Furthermore, γ upperbounds the log-ratio of any two functions $\phi_\theta(x), \phi_{\theta'}(x)$ for all θ, θ', x and therefore

$$\sup_{\theta, \theta', x} \log \frac{\phi_\theta(x)}{\phi_{\theta'}(x)} < \infty \tag{2}$$

is a condition on the class $\mathcal{H}$.

Li and Barron prove that k-mixture approximations satisfying (1) can be constructed by the following greedy procedure: Initialize $f_1 = \phi_\theta$ to minimize $D(f\|f_1)$ and at step k construct f_k from f_{k-1} by finding α and θ such that

$$D(f\|f_k) \leq \min_{\alpha, \theta} D(f\|(1-\alpha) f_{k-1}(x) + \alpha \phi_\theta(x)).$$

Furthermore, a connection between KL-divergence and Maximum Likelihood suggests the following method to compute the *estimate* $\hat{f}_k$ *from the data* by greedily choosing ϕ_θ at step k so that

$$\sum_{i=1}^{n} \log \hat{f}_k(x_i) \geq \max_{\alpha,\theta} \sum_{i=1}^{n} \log[(1-\alpha)\hat{f}_{k-1}(x_i) + \alpha\phi_\theta(x_i)] \tag{3}$$

Li and Barron proved the following theorem:

Theorem 1.1. *Let $\hat{f}_k(x)$ be either the maximizer of the likelihood over k-component mixtures or more generally any sequence of density estimates satisfying (3). Assume additionally that Θ is a d-dimensional cube with side-length A, and that*

$$\sup_{x\in\mathcal{X}} |\log \phi_\theta(x) - \log \phi_{\theta'}(x)| \leq B \sum_{j}^{d} |\theta_j - \theta'_j| \tag{4}$$

for any $\theta, \theta' \in \Theta$. Then

$$\mathbb{E}_S\left[D(f\|\hat{f}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2 k}{n}\log(nc_3), \tag{5}$$

where c_1, c_2, c_3 are constants (dependent on A, B, d).

Here $\mathbb{E}_S$ denotes the expectation with respect to a draw of n independent points according to the unknown distribution f. The above bound combines the *approximation* and *estimation* results. Note that the first term decreases with the number of components k, while the second term increases. The rate of convergence for the optimal k is therefore $O(\sqrt{\frac{\log n}{n}})$.

2 Main Results

Instead of condition (2), we assume that class $\mathcal{H}$ consists of functions bounded above and below by a and b, respectively. See the discussion section for the comparison of these two assumptions. We prove the following results:

Theorem 2.1. *For any target density f such that $a \leq f \leq b$ and $\hat{f}_k(x)$ either the maximizer of the likelihood over k-component mixtures or more generally any sequence of density estimates satisfying (3),*

$$\mathbb{E}_S\left[D(f\|\hat{f}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}_S\left[\frac{c_2}{\sqrt{n}}\int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_x)d\epsilon\right].$$

where c_1, c_2 are constants (dependent on a, b) and $\mathcal{D}(\mathcal{H}, \epsilon, d_x)$ is the covering number of $\mathcal{H}$ at scale ϵ with respect to empirical distance d_x.

Corollary 2.1. *Under the conditions of Theorem 1.1 (i.e. $\mathcal{H}$ satisfying condition (4) and Θ being a cube with side-length A), the bound of Theorem 2.1 becomes*

$$\mathbb{E}_S\left[D(f\|\hat{f}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}},$$

where c_1 and c_2 are constants (dependent on a, b, A, B, d).

3 Discussion of the Results

The result of Theorem 2.1 is twofold. The first implication concerns dependence of the bound on k, the number of components. Our results show that there is an estimation bound of the order $O(\frac{1}{\sqrt{n}})$ that does not depend on k. Therefore, the number of components is not a trade-off that has to be made with the approximation part.
The second implication concerns the rate of convergence in terms of n, the number of samples. The rate of convergence (in the sense of KL-divergence) of the estimated mixture to the true density is of the order $O(1/\sqrt{n})$. As Corollary 2.1 shows, for the specific class $\mathcal{H}$ considered by Li and Barron, the Dudley integral converges and does not depend on n. Furthermore, the result of this paper holds for general base classes $\mathcal{H}$ with a converging entropy integral, extending and improving the result of Li and Barron. Note that the bound of Theorem 2.1 is in terms of the metric entropy of $\mathcal{H}$, as opposed to the metric entropy of $\mathcal{C}$. This is a strong result because the convex class $\mathcal{C}$ can be very large [8] even for small $\mathcal{H}$.
Rates of convergence for the MLE in mixture models were recently studied by Sara van de Geer [10]. As the author notes, the optimality of the rates depends primarily on the optimality of the entropy calculations. Unfortunately, in the results of [10], the entropy of the convex class appears in the bounds, which is undesirable. Moreover, only finite combinations are considered. Wong and Shen [13] also considered density estimation, giving rates of convergence in Hellinger distance for a class of bounded Lipschitz densities. In their work, again, a bound on the metric entropy of the whole class is used and the rates of convergence are slower than those achieved in this paper.
An advantage of the approach of [10] is the use of Hellinger distance to avoid problems near zero. Li and Barron address this problem by requiring (2), which is boundedness of the log of the ratio of two densities. We address this problem by assuming boundedness of the densities directly. The two conditions are equivalent unless we consider classes consisting only of unbounded functions or consisting only of functions approaching 0 at the same rate (in which case condition (2) is weaker). If the boundedness of densities is assumed, as [3] notes, the KL-divergence and the Hellinger distance do not differ by more than a multiplicative constant.

4 Proofs

Assume $0 < a \leq \phi_\theta \leq b$ for all $\phi_\theta \in \mathcal{H}$. Constants which depend only on a and b we will denote by c with various subscripts. The values of the constants might change from line to line.

Theorem 4.1. *For any fixed f, $0 < a \leq f \leq b$ and $S = \{x_1, ..., x_n\}$ drawn i.i.d from f, with probability at least $1 - e^{-t}$,*

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \mathbb{E}_S \left[\frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_x) d\epsilon \right] + c_2 \sqrt{\frac{t}{n}}$$

where c_1 and c_2 are constants that depend on a and b.

Proof By Lemma A.3,

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$ and by Lemma A.2,

$$\mathbb{E}_S \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq 2\mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right|.$$

Combining,

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq 2\mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right| + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$.
Therefore, instead of bounding the difference between the "empirical" and the "expectation", it is enough to bound the above expectation of the Rademacher average. This is a simpler task, but first we have to deal with the log and the fraction (over f) in the Rademacher sum. To eliminate these difficulties, we apply Lemma A.1 twice. Once we reduce our problem to bounding the Rademacher sum $\sup_{\phi \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(x_i) \right|$ of the basis functions, we will be able to use the entropy of the class $\mathcal{H}$.
Let $p_i = \frac{h(x_i)}{f(x_i)} - 1$. and note that $\frac{a}{b} - 1 \leq p_i \leq \frac{b}{a} - 1$. Consider $\phi(p_i) = \log(1 + p_i)$. The largest derivative of $\log(1+p)$ on the interval $p \in [\frac{a}{b} - 1, \frac{b}{a} - 1]$ is at $p = a/b - 1$ and is equal to b/a. So, $\frac{a}{b} \log(p+1)$ is 1-Lipschitz. Also, $\phi(0) = 0$. By Lemma A.1 applied to $\phi(p_i)$ and G being identity mapping,

$$\begin{aligned}
2\mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \frac{h(x_i)}{f(x_i)} \right| &= 2\mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{1}^{n} \epsilon_i \phi(p_i) \right| \\
&\leq 2\frac{b}{a} \mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} - \frac{1}{n} \sum_{1}^{n} \epsilon_i \right| \\
&\leq 2\frac{b}{a} \mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} \right| + 2\frac{b}{a} \mathbb{E}_{\epsilon} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right| \\
&\leq 2\frac{b}{a} \mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} \right| + 2\frac{b}{a} \frac{1}{\sqrt{n}}.
\end{aligned}$$

The last inequality holds trivially by upperbounding L_1 norm by the L_2 norm. Now apply A.1 again with contraction $\phi_i(h_i) = a\frac{h_i}{f_i}$.

$$|\phi_i(h_i) - \phi_i(g_i)| = \frac{a}{|f_i|} |h_i - g_i| \leq |h_i - g_i|$$

$$2\frac{b}{a} \mathbb{E}_{S,\epsilon} \sup_{fh \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{h(x_i)}{f(x_i)} \right| \leq 2\frac{b}{a^2} \mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right|.$$

Combining the inequalities, with probability at least $1 - e^{-t}$

$$\sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - \mathbb{E} \log \frac{h}{f} \right| \leq \frac{2b}{a^2} \mathbb{E}_{S,\epsilon} \sup_{h \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i h(x_i) \right| + \sqrt{8} \log \frac{b}{a} \sqrt{\frac{t}{n}} + \frac{2b}{a} \frac{1}{\sqrt{n}}.$$

The power of using Rademacher averages to estimate complexity comes from the fact that the Rademacher averages of a class are equal to those of the convex hull. Indeed, consider $\sup_{h\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i)\right|$ with $h(x) = \int_\theta \phi_\theta(x) P(d\theta)$. Since a linear functional of convex combinations achieves its maximum value at the vertices, the above supremum is equal to

$$\sup_\theta \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \phi_\theta(x_i)\right|,$$

the corresponding supremum on the basis functions ϕ. Therefore,

$$\mathbb{E}_\epsilon \sup_{h\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i)\right| = \mathbb{E}_\epsilon \sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \phi_\theta(x_i)\right|.$$

Next, we use the following classical result [12],

$$\mathbb{E}_\epsilon \sup_{\phi\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \phi(x_i)\right| \leq \frac{c_1}{\sqrt{n}} \int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_x) d\epsilon,$$

where d_x is the empirical distance with respect to the set S.
Putting it all together, the following holds with probability at least $1 - e^{-t}$:

$$\sup_{h\in\mathcal{C}} \left|\frac{1}{n}\sum_{i=1}^n \log\frac{h(x_i)}{f(x_i)} - \mathbb{E}\log\frac{h}{f}\right| \;\leq\; \mathbb{E}_S\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right] + c_2\sqrt{\frac{t}{n}}.$$

If $\mathcal{H}$ is a VC-subgraph with VC dimension V, the Dudley integral above is bounded by $c\sqrt{V}$ and we get $\frac{1}{\sqrt{n}}$ convergence. One example of such a class is worked out in the Appendix (Gaussian densities over a bounded domain and with bounded variance). Another example is the class considered in [6], and the cover is computed for it in the proof of Corollary 2.1. □

We are now ready to prove Theorem 2.1:

Proof

$$\begin{aligned}
D(f\|\hat{f}_k) - D(f\|f_k) &= \left(\mathbb{E}\log\frac{f}{\hat{f}_k} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\hat{f}_k(x_i)}\right) + \left(\frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{f_k(x_i)} - \mathbb{E}\log\frac{f}{f_k}\right) \\
&+ \left(\frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\hat{f}_k(x_i)} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{f_k(x_i)}\right) \\
&\leq 2\sup_{h\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^n \log\frac{h(x_i)}{f(x_i)} - \mathbb{E}\log\frac{h}{f}\right| \\
&+ \left(\frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\hat{f}_k(x_i)} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{f_k(x_i)}\right) \\
&\leq \mathbb{E}_S\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right] + c_2\sqrt{\frac{t}{n}} + \frac{1}{n}\sum_{i=1}^n \log\frac{f_k(x_i)}{\hat{f}_k(x_i)}
\end{aligned}$$

with probability at least $1 - e^{-t}$ (by Theorem 4.1). Note that $\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_k(x_i)}{\hat{f}_k(x_i)} \leq 0$ if $\hat{f}_k$ is constructed by maximizing likelihood over k-component mixtures. If it is constructed by a greedy algoritheorem described in the previous section, $\hat{f}_k$ achieves "almost maximum likelihood" ([7]) in following sense:

$$\forall g \in \mathcal{C}, \quad \frac{1}{n}\sum_{i=1}^{n} \log(\hat{f}_k(x_i)) \geq \frac{1}{n}\sum_{i=1}^{n} \log(g(x_i)) - \gamma \frac{c^2_{F_n,P}}{k}.$$

Here $c^2_{F_n,P} = (1/n)\sum_{i=1}^{n} \frac{\int \phi^2_\theta(x_i)P(d\theta)}{\left(\int \phi_\theta(x_i)P(d\theta)\right)^2} \leq \frac{b^2}{a^2}$ and $\gamma = 4\log(3\sqrt{e}) + 4\log\frac{b}{a}$. Hence, with probability at least $1 - e^{-t}$,

$$D(f\|\hat{f}_k) - D(f\|f_k) \leq \mathbb{E}_S\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right] + c_2\sqrt{\frac{t}{n}} + \frac{c_3}{k}.$$

We now write the overall error of estimating an unknown density f as the sum of approximation and estimation errors. The former is bounded by (1) and the latter is bounded as above. Note again that $c^2_{f,P}$ and γ in the approximation bound (1) are bounded above by constants which depend only on a and b. Therefore, with probability at least $1 - e^{-t}$,

$$\begin{aligned} D(f\|\hat{f}_k) - D(f\|\mathcal{C}) &= (D(f\|f_k) - D(f\|\mathcal{C})) + \left(D(f\|\hat{f}_k) - D(f\|f_k)\right) \\ &\leq \frac{c}{k} + \mathbb{E}_S\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right] + c_2\sqrt{\frac{t}{n}}. \end{aligned}$$

Finally, we rewrite the above probabilistic statement as a statement in terms of expectations. Let $\zeta = \frac{c}{k} + \mathbb{E}_S\left[\frac{c_1}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right]$ and $\xi = D(f\|\hat{f}_k) - D(f\|\mathcal{C})$. We have shown that

$$\mathbb{P}\left(\xi \geq \zeta + c_2\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

Since $\xi \geq 0$,

$$\begin{aligned} \mathbb{E}_S[\xi] &= \int_0^\infty \mathbb{P}(\xi > u)\,du = \int_0^\zeta \mathbb{P}(\xi > u)\,du + \int_\zeta^\infty \mathbb{P}(\xi > u)du \\ &\leq \zeta + \int_0^\infty \mathbb{P}(\xi > u + \zeta)\,du. \end{aligned}$$

Now set $u = c_2\sqrt{\frac{t}{n}}$. Then $t = c_3 n u^2$ and

$$E_S[\xi] \leq \zeta + \int_0^\infty e^{-c_3 n u^2}du \leq \zeta + \frac{c}{\sqrt{n}}.$$

Hence,

$$E_S\left[D(f\|\hat{f}_k)\right] - D(f\|\mathcal{C}) \leq \frac{c_1}{k} + \mathbb{E}_S\left[\frac{c_2}{\sqrt{n}}\int_0^b \log^{1/2}\mathcal{D}(\mathcal{H},\epsilon,d_x)d\epsilon\right].$$

□

Remark 4.1. *In the actual proof of the bounds, Li and Barron [7, 6] use a specific sequence of α_i for the finite combinations. The authors take $\alpha_1 = 1$, $\alpha_2 = \frac{1}{2}$, and $\alpha_k = \frac{2}{k}$ for $k \geq 2$. It can be shown that with these weights*

$$f_k = \frac{2}{k(k-1)} \left(\frac{1}{2}\phi_1 + \frac{1}{2}\phi_2 + \sum_{m=3}^{k} (m-1)\phi_m \right),$$

so the later choices have more weight.

We now prove Corollary 2.1:

Proof Since we consider bounded densities $a \leq \phi_\theta \leq b$, condition (4) implies that

$$\forall x,\ \log\left(\frac{\phi_\theta(x) - \phi_{\theta'}(x)}{b} + 1 \right) \leq B|\theta - \theta'|_{L_1}.$$

This allows to bound L_∞ distances between functions in $\mathcal{H}$ in terms of the L_1 distances between the corresponding parameters. Since Θ is a d-dimensional cube of side-length A, we can cover Θ by $\left(\frac{A}{\delta}\right)^d$ "balls" of L_1-radius $d\frac{\delta}{2}$. This cover induces a cover of $\mathcal{H}$. For any f_θ there exists an element of the cover $f_{\theta'}$, so that the

$$d_x(f_\theta, f_{\theta'}) \leq |f_\theta - f_{\theta'}|_\infty \leq b e^{B\frac{d\delta}{2}} - b = \epsilon.$$

Therefore, $\delta = \frac{2\log\left(\frac{\epsilon}{b}+1\right)}{Bd}$ and the cardinality of the cover is $(\frac{A}{\delta})^d = \left(\frac{ABd}{2\log\left(\frac{\epsilon}{b}+1\right)} \right)^d$. So,

$$\int_0^b \log^{1/2} \mathcal{D}(\mathcal{H}, \epsilon, d_x) d\epsilon = \int_0^b \sqrt{d \log \frac{ABd}{2\log\left(\frac{\epsilon}{b}+1\right)}} d\epsilon.$$

A straightforward calculation shows that the integral above converges. □

5 Future Work

The main drawback of the approach described in this paper is the need to lower-bound the densities. Future work will focus on ways to remove this condition by using, for instance, a truncation argument.

A Appendix

We will denote $f_i = f(x_i)$. The following inequality can be found in [5], Theorem 4.12.

Lemma A.1 ([5] Comparison inequality for Rademacher processes). *If $G : \mathbb{R} \to \mathbb{R}$ convex and non-decreasing and $\phi_i : \mathbb{R} \to \mathbb{R}$ $(i = 1, .., n)$ contractions ($\phi_i(0) = 0$ and $|\phi_i(s) - \phi_i(t)| \leq |s - t|$), then*

$$\mathbb{E}_\epsilon G(\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i \phi_i(f_i)) \leq \mathbb{E}_\epsilon G(\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i f_i).$$

Lemma A.2 ([12] Symmetrization). *Consider the following processes:*

$$Z(x) = \sup_{f\in\mathcal{F}} \left| \mathbb{E}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right|, R(x) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i) \right|.$$

Then

$$\mathbb{E}Z(x) \leq 2\mathbb{E}R(x).$$

Lemma A.3 (Application of McDiarmid inequality). *For*

$$Z(x_1, ..., x_n) = \sup_{h\in\mathcal{F}} \left| \mathbb{E}\log\frac{h}{f} - \frac{1}{n}\sum_{i=1}^{n} \log\frac{h(x_i)}{f(x_i)} \right|$$

the following holds with probability at least $1 - e^{-t}$*:*

$$Z - \mathbb{E}Z \leq c\sqrt{\frac{t}{n}},$$

where a *and* b *are the lower and upper bounds for* f *and* h*, and* $c = 2\sqrt{2}\log\frac{b}{a}$*.*

Proof Let $t_i = \log\frac{h(x_i)}{f(x_i)}$ and $t_i' = \log\frac{h(x_i')}{f(x_i')}$. The bound on the martingale difference follows:

$$\begin{aligned} |Z(x_1, ..., x_i', ..., x_n) - Z(x_1, ..., x_i, ..., x_n)| &= \\ \left| \sup_{h\in\mathcal{F}} \left| \mathbb{E}\log\frac{h}{f} - \frac{1}{n}(t_1 + ... + t_i + ... + t_n) \right| - \sup_{h\in\mathcal{F}} \left| \mathbb{E}\log\frac{h}{f} - \frac{1}{n}(t_1 + ... + t_i' + ... + t_n) \right| \right| &\leq \\ \leq \sup_{h\in\mathcal{F}} \frac{1}{n} \left| \log\frac{h(x_i')}{f(x_i')} - \log\frac{h(x_i)}{f(x_i)} \right| \leq \frac{1}{n}\left(\log\frac{b}{a} - \log\frac{a}{b} \right) = \frac{1}{n} 2\log\frac{b}{a} &= c_i. \end{aligned}$$

The above chain of inequalities holds because of triangle inequality and properties of sup. Applying McDiarmid's inequality,

$$\mathbb{P}\left(Z - \mathbb{E}Z > u\right) \leq \exp\left(-\frac{u^2}{2\sum c_i^2} \right) = \exp\left(-\frac{nu^2}{8\log^2\frac{b}{a}} \right).$$

Equivalently,

$$\mathbb{P}\left(Z - \mathbb{E}Z > c\sqrt{\frac{t}{n}} \right) \leq e^{-t},$$

for constant $c = 2\sqrt{2}\log\frac{b}{a}$. □

B Example of Gaussian Densities

Let $\mathcal{F} = \{f_{\mu,\sigma} : f_{\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), |\mu| \leq M, \sigma_{min} \leq \sigma \leq \sigma_{max}\}$ be a set of Gaussian densities defined over a bounded set $\mathcal{X} = [-M, M]$ with bounded variance. Here we show that $\mathcal{F}$ has a finite cover $\mathcal{D}(\mathcal{F}, \epsilon, d_x) = \frac{K}{\epsilon^2}$, for some constant K.

Define

$$\mathcal{F}_\mu = \{f_{\mu,\sigma} : f_{\mu,\sigma} \in \mathcal{F}, \mu \in \{-M + k\epsilon_\mu : k = 0, ..., 2M/\epsilon_\mu\}\}$$

and

$$\mathcal{F}_{\mu,\sigma} = \{f_{\mu,\sigma} : f_{\mu,\sigma} \in \mathcal{F}_\mu, \sigma \in \{\sigma_{min} + k\epsilon_\sigma : k = 0, ..., (\sigma_{max} - \sigma_{min})/\epsilon_\sigma\}\}.$$

Thus, $\mathcal{F}_{\mu,\sigma} \subset \mathcal{F}_\mu \subset \mathcal{F}$. We claim that $\mathcal{F}_{\mu,\sigma}$ is finite ϵ-cover for $\mathcal{F}$ with respect to the d_x norm (on the data). For any $f_{\mu,\sigma} \in \mathcal{F}$, first choose a function $f_{\mu',\sigma} \in \mathcal{F}_\mu$ so that $|\mu - \mu'| \leq \epsilon_\mu$. Note that functions $f \in \mathcal{F}$ are all Lipschitz because σ is bounded. In fact, largest derivative of $f = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ is at $-\sigma$ and is equal to $\frac{1}{\sqrt{2\pi e}\sigma^2}$. Then

$$|f_{\mu,\sigma}(x) - f_{\mu',\sigma}(x)| \leq \frac{1}{\sqrt{2\pi e}\sigma^2}|\mu - \mu'| \leq \frac{\epsilon_\mu}{\sqrt{2\pi e}\sigma_{min}^2}.$$

Furthermore, any $f_{\mu',\sigma} \in \mathcal{F}_\mu$ can be approximated by $f_{\mu',\sigma'} \in \mathcal{F}_{\mu,\sigma}$ such that $|\sigma - \sigma'| \leq \epsilon_\sigma$. Then

$$\forall x \in \mathcal{X} \quad |f_{\mu',\sigma}(x) - f_{\mu',\sigma'}(x)| \leq \frac{1}{\sqrt{2\pi}} \left|\frac{1}{\sigma} - \frac{1}{\sigma'}\right| \leq \frac{1}{\sqrt{2\pi}} \frac{\epsilon_\sigma}{\sigma_{min}^2}.$$

Combining the two steps, any function in $\mathcal{F}$ can be approximated by a function in $\mathcal{F}$ with an error at most $(\epsilon_\mu + \epsilon_\sigma)\frac{1}{\sigma_{min}^2\sqrt{2\pi}}$. The empirical distance

$$\begin{aligned} d_x(f_{\mu,\sigma}, f_{\mu',\sigma'}) &= \left(\frac{1}{n}\sum_{i=1}^{n}(f_{\mu,\sigma}(x) - f_{\mu',\sigma'}(x))^2)\right)^{\frac{1}{2}} \leq \sup_x |f_{\mu,\sigma}(x) - f_{\mu',\sigma'}(x)| \\ &\leq (\epsilon_\mu + \epsilon_\sigma)\frac{1}{\sigma_{min}^2\sqrt{2\pi}} = \epsilon. \end{aligned}$$

Choosing $\epsilon_\mu = \epsilon_\sigma = \epsilon\frac{\sigma_{min}^2\sqrt{\pi}}{\sqrt{2}}$ we get the size of the cover to be

$$\mathcal{D}(\mathcal{F}, \epsilon, d_x) = card(\mathcal{F}_{\mu,\epsilon}) = \frac{2M}{\epsilon_\mu}\frac{(\sigma_{max} - \sigma_{min})}{\epsilon_\sigma} = \frac{4M(\sigma_{max} - \sigma_{min})}{\pi\sigma_{min}^4}\frac{1}{\epsilon^2} = \frac{K}{\epsilon^2}.$$

References

[1] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transaction on Information Theory*, 39(3):930–945, May 1993.

[2] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

[3] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[4] L.K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for Projection Pursuit Regression and neural network training. *The Annals of Statistics*, 20(1):608–613, March 1992.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[6] J. Li and A. Barron. Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural information processings systems 12*, San Mateo, CA, 1999. Morgan Kaufmann Publishers.

[7] Jonathan Q. Li. *Estimation of Mixture Models*. PhD thesis, The Department of Statistics. Yale University, 1999.

[8] Shahar Mendelson. On the size of convex hulls of small sets. *Journal of Machine Learning Research*, 2:1–18, 2001.

[9] P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics*, 10:51–80, 1999.

[10] S.A. van de Geer. Rates of convergence for the maximum likelihood estimator in mixture models. *Nonparametric Statistics*, 6:293–310, 1996.

[11] S.A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[12] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.

[13] W.H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates for sieve mles. *Annals of Statistics*, 23:339–362, 1995.